



On the Front Lines of Biodefense

*A Livermore team is developing
DNA-based signatures to quickly
and accurately identify pathogens.*

WHEN a sudden outbreak of a strange virus called Severe Acute Respiratory Syndrome (SARS) occurred last year, the Centers for Disease Control and Prevention (CDC) sought help from a team of Lawrence Livermore biologists, mathematicians, and computer scientists. Within three hours of receiving the first sequenced genome (genetic blueprint) of the virus from the CDC, the Livermore team produced several candidate signatures of the pathogen (disease-causing microbe). Signatures are specific regions of DNA or RNA that uniquely identify a pathogen. The SARS case was one of many in which the group has developed signatures using a novel whole-genome analysis approach that is changing pathogen diagnostic design.

Sensors Sniff City Air for Pathogens

The specter of terrorists attacking American cities with airborne pathogens has prompted federal agencies to develop systems that continuously monitor the air for biothreat agents.

The nation's first such monitoring system is the Biological Aerosol Sentry and Information System (BASIS). BASIS was developed by a team of researchers from Lawrence Livermore and Los Alamos national laboratories and involved extensive collaborations with emergency response, public health, and law-enforcement agencies. The system uses detection methods derived from DNA-based signatures designed by Livermore's bioinformatics group. BASIS was designed for the "detect to treat" mission, identifying a release quickly enough to permit effective medical treatment of those exposed.

BASIS was called into service after the anthrax attacks of October 2001. It was also deployed to Salt Lake City, Utah, as part of the overall security strategy for the 2002 Winter Olympic Games. (See *S&TR*, October 2003, pp. 6–7.) The system was later deployed in Albuquerque during the summer of 2002 and in New York City for the first anniversary of the September 11 terrorist attacks.

BASIS air-monitoring units collect aerosol samples at specific locations. A semi-automated mobile field laboratory rapidly analyzes DNA from the collected samples for evidence of potentially lethal bacteria and viruses. Safeguards built into the system ensure sample integrity. Should a positive identification be confirmed, the field laboratory immediately notifies the appropriate response agencies.

In late 2002, the U.S. Department of Homeland Security, the Environmental Protection Agency, and the Centers for Disease Control and Prevention implemented the national BioWatch program.

Some BioWatch sensors resemble a phone booth topped with an air intake and radio antenna. Couriers collect air filters from the sensors and deliver them to military facilities or public health laboratories. There, technicians use Livermore-developed signatures to detect the presence of target pathogens. If a pathogen were detected, officials would examine wind patterns in the area of the contaminated sensor and take action to protect the population.

In summer 2003, BioWatch sensors in Houston detected fragments of *Francisella tularensis*, a bacterium found in rabbits, prairie dogs, and rodents that can spread to humans and cause tularemia. Health officials concluded there had been no attack. Instead, the sensor had detected tiny amounts of *F. tularensis* naturally present in the environment. Although *F. tularensis* was known to be endemic in Texas, this was the first time it was detected in an aerosol sample.

The Department of Homeland Security announced that the incident marked the first time the BioWatch network had detected such a serious airborne threat, one that in this case was naturally occurring. Tom Slezak, leader of Livermore's bioinformatics group, notes that in more than 700,000 uses of Livermore-developed signatures, the BASIS/BioWatch network of sensors has never raised a false-positive alarm, that is, concluded that pathogens were present when they were not.



(a) Air samples are collected by this BASIS sensor installed outside a New York Police station. (b) Another sensor resides in a New York City borough neighborhood. (c) New York City Department of Public Health officers accompany a Department of Energy employee retrieving air samples from a sensor located near the former World Trade Center. Air samples are tested by Livermore technicians at a nearby laboratory.

The team, part of the Laboratory's Biology and Biotechnology Research Program (BBRP), has been on the front lines of the nation's biodefense effort since 2001. Eleven computer scientists, biologists, and mathematicians led by computer scientist Tom Slezak comprise one of the largest pathogen bioinformatics groups. Their work spans the full spectrum of effort, from identifying signature candidates to developing DNA-based signatures and deploying validated assays in the field. Team members have traveled throughout the nation, often with only a few hours' notice, to support the national effort to defend against bioterrorism.

Biological weapons could include bacteria (anthrax, plague), DNA viruses (smallpox), RNA viruses (ebola, SARS, foot-and-mouth disease), fungi (soybean rust, corn rust), protozoa (giardiasis), and toxins (ricin). Pathogens such as these and many others could be used to sicken or kill urban populations, livestock, or crops. Early detection and unmistakable identification are crucial to limiting the potentially catastrophic human and economic costs of a bioattack.

Many types of signature requests are received by the team. One request may be for all strains of a normally pathogenic species, including its nonpathogenic and vaccine strains. Another request may be for all of the pathogenic strains of a particular

species. Fulfilling these requests can be difficult because while there may be hundreds of strains of a particular species, genomic sequences may exist for only a few. Strains may also vary in pathogenicity, and their genetic near-neighbors may or may not be virulent or may affect hosts other than humans. In addition, RNA viruses have extremely high mutation rates, so it may be difficult or impossible to find adequate stable regions suitable for use as a signature.

The Livermore bioinformatics team has developed DNA-based signatures of virtually every biothreat pathogen (the organisms identified by the CDC as high-priority threat agents) for which adequate genomic sequences are available as well as for several other human and livestock pathogens. Signature requests come from agencies such as the Department of Energy (DOE), the CDC's Laboratory Response Network and BioWatch Program, the Department of Agriculture, the Food and Drug Administration, and the Department of Defense. Livermore signatures are part of the nation's public health system and have been in use for homeland defense since fall 2001. (See the [box on p. 5.](#))

Pipeline Called KPATH

Livermore's signature pipeline, called KPATH, is used to develop the signatures of bacterial and viral pathogens. This

Livermore-designed system is a fully automated DNA-based signature "pipeline," able to deliver signature candidates (spanning 200–300 base pairs of DNA) in minutes to hours. In simplest terms, KPATH works by comparing the genome of the target pathogen to a library of microbial genomes, searching for those areas that are unique to the target organism. (See the [box on p. 8.](#))

KPATH uses the software programs Multiple Genome Aligner (MGA) and Vmatch, which were developed by collaborators in Germany. MGA aligns the multiple genomes of a target pathogen, and Vmatch uses efficient algorithms to quickly compare the genome of interest with all other sequenced microbial genomes. "These software tools allow the pathogen genomes themselves to show us which regions of DNA are important," Slezak says. The DNA regions that are significant to the pathogen are conserved among all strains of the pathogen sequenced to date and are unique when compared to all other organisms sequenced to date. That is, they are present in every strain of the pathogen and absent in all other organisms.

The algorithms work by locating those portions of the genome that are not unique and eliminating them from consideration. "In this way," says Slezak, "we define regions of apparent uniqueness and mine them for candidate signatures."

Candidate signatures must then be verified in the laboratory. "It's a long path from candidate signature to validated assay," notes Slezak. Hundreds of thousands of candidate signatures are computationally screened. Wet-chemistry procedures reduce that number to hundreds and then dozens. Much of the laboratory testing takes place at the CDC and other organizations that are certified to work with virulent pathogens. Once a signature is verified, the final step is optimizing the signature for a specific detection chemistry or instrument using a specific protocol. When that process is complete, the signature is called an assay.

One of KPATH's important features is that it automatically downloads newly sequenced pathogen genomes from all major public

Livermore researchers try to sleep on a U.S. Air Force C-130 transport plane on their way to deploy a pathogen-detection system.



databases, and all validated and fielded assays are verified weekly as the new sequence data are acquired. “As known strains evolve and new strains are discovered and their genomes sequenced, some of the ‘unique’ regions will erode,” says Slezak. “We’ll then need to refine the signature.”

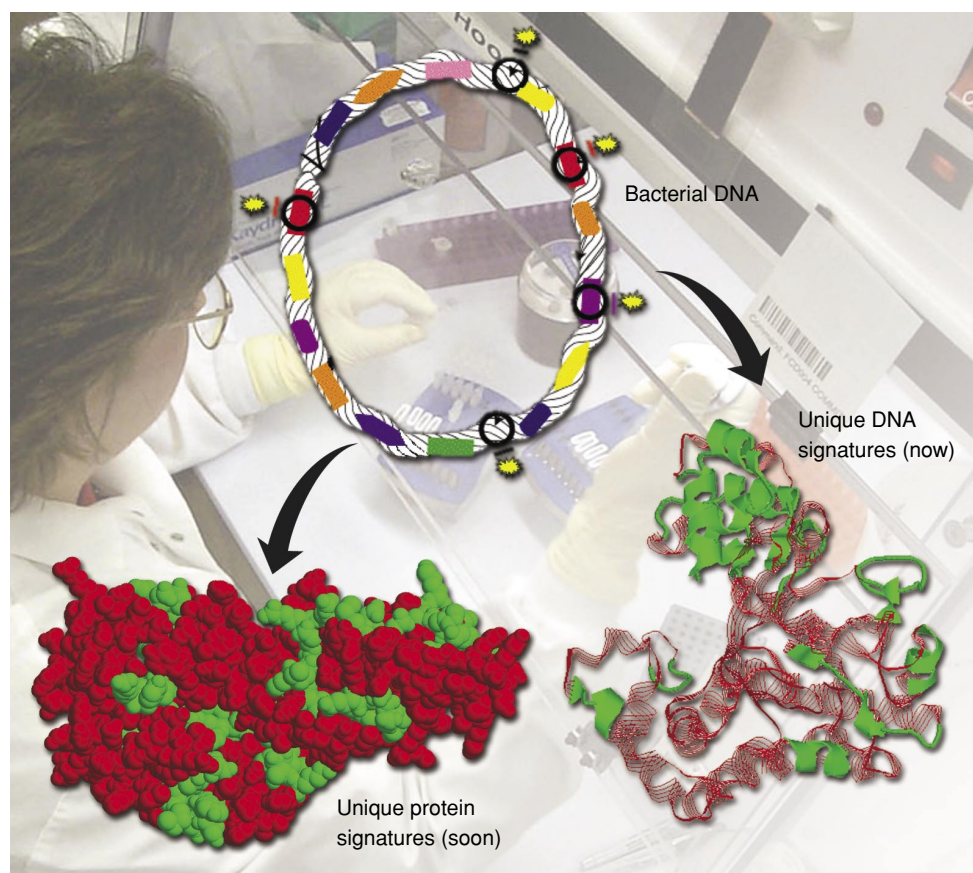
Olympic Games Motivate

In early 2000, the DOE’s Chemical and Biological National Security Program (CBNP) began a national pathogen-detection effort following the announcement by then-Secretary Richardson that DOE would be providing biosecurity at the 2002 Winter Olympic Games in Salt Lake City, Utah. Lawrence Livermore was assigned the task of developing reliable and validated assays for a number of the most likely bioterrorism agents. (See the [box on p. 5.](#))

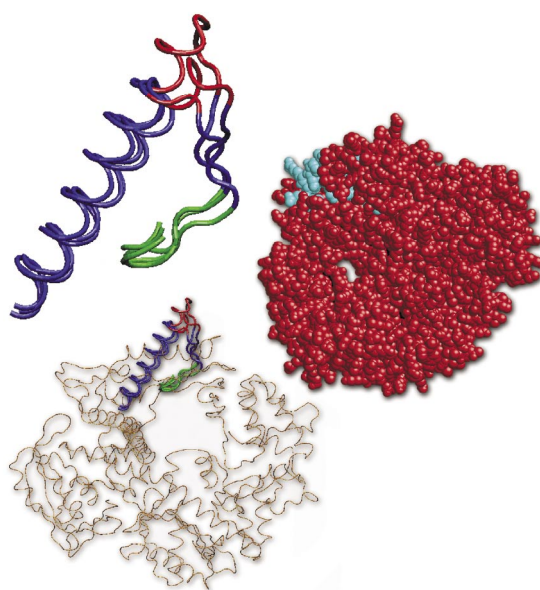
The bioinformatics team reasoned that a whole-genome analysis approach—that is, comparing a target pathogen genome against all other sequenced microbial genomes—would reveal which regions of the DNA were unique. They also believed the process could be automated to get results more quickly. Until the Livermore approach, signature design was a time-consuming, expensive process done largely by hand and guided heavily by intuition. Analysis was generally limited to sequences from a few genes thought to be important. Traditional approaches to DNA-based signature development started with the assumption that a particular gene was vital to an organism’s virulence, host range, or other factor. The resulting assay would then be tested with the available strain. This approach would at times yield good results, but it frequently resulted in failure.

Computational support for diagnostic development was rare. “The time was ripe for radical changes in this field,” says Slezak.

In August 2000, the team began building a set of tools that would accomplish these goals. Slezak says, “We used techniques and mindsets from our many years of experience working on the Human Genome Project (HGP).” Slezak formerly led Livermore’s HGP bioinformatics effort and later the



There are two major ways to detect pathogen signatures. One involves finding specific regions of DNA (or RNA) that uniquely identify a pathogen. The other (still under development) involves finding specific regions of a unique protein whose production is specified by that pathogen’s DNA (or RNA).



DNA-based signatures are often the part of a pathogen’s genome that reflects the code of a particular protein or enzyme unique to the pathogen. This model depicts one such pathogen protein. Blue represents the part of the protein that is conserved (present) in all strains of the pathogen’s DNA. Red represents the portion of the protein that is unique to the target strain. Green depicts the most highly conserved portion across multiple organisms. The DNA-based signature of this target strain, therefore, combines those portions of DNA that code for both the blue and red regions.

On the Road to KPATH: A Short Timeline

In 2000, a crude test on *Bacillus anthracis*, the causative agent of anthrax, demonstrated that a computer-based approach to pathogen-signature development would work. The test took summer student Marisa Lam several days to process more than 4 billion bytes of information from analysis of the *B. anthracis* genome. The roughly 4,000 resulting candidate signatures were narrowed down to a handful and forwarded to the Centers for Disease Control and Prevention (CDC) for formal validation. These optimized signatures became the assays used for the nation's Biological Aerosol Sentry and Information System (BASIS) and BioWatch environmental monitoring networks. By comparison, earlier methods of signature design had yielded zero successes among more than 1,000 candidates. Buoyed by this successful test, a team led by computer scientist Tom Kuczmarski began work on an automated signature pipeline.

In February 2001, foot-and-mouth disease was devastating the cattle and sheep industry in the United Kingdom. The Livermore team analyzed the tiny (8-kilobase) genome of the foot-and-mouth disease virus (FMDV) and found that viral genomes, although tiny compared to the typically 3- to 5-megabase bacterial genomes, can be troublesome to analyze because of their high mutation rate. The team determined that only one region of the FMDV genome is capable of supporting a signature assay. In the spring, the team began collaborating with Sharon Hietala of the University of California at Davis to detect agricultural pathogens that are common to California cattle and cause symptoms that mimic those of FMDV.

Livermore computer scientist Tom Slezak was attending a conference in Maryland when the September 11 terrorist attacks occurred. During the five days it took him to return to California, he conceived of a fully automated DNA-based signature-design-and-maintenance system that would download all new and updated genomic sequences weekly from the major public databases. The system would then compare those sequences with the existing, fielded DNA-based signatures to determine if any new sequences had invalidated them. Slezak named the system KPATH. "I was inspired by radio-station call signals," says Slezak. "KPATH, all pathogens, all the time. Bringing you the pathogen hits of the 50s, 60s, and today." To achieve the desired speed and capacity, KPATH would require a large server with multiple central-processing units (CPU), a powerful database server, and more advanced algorithms.



The Livermore signature development team includes (from left) Nisha Mulakken, Carol Zhou, Adam Zemla, Ed Miller, Tom Slezak, Tom Kuczmarski, Jason Smith, Marisa Lam, Clinton Torres, Shea Gardner, Mark Wagner, and Beth Vitalis. The Livermore team collectively has expertise in biology, mathematics, systems science, and computer science.

In October 2001, when the anthrax attacks occurred, BASIS was the only system in the country capable of taking on pathogen monitoring duties. "The anthrax attacks resulted in the first real awareness of bioterrorism by most of the U.S. general public," notes Slezak. A few months later, BASIS also took on monitoring duties at the 2002 Winter Olympic Games.

In June 2002, supplemental funding was obtained from the Department of Energy to purchase a 24-CPU server, an 8-CPU database machine, and a 3-terabyte file server and to hire six summer students to help build KPATH. The team also adopted the new Multiple Genome Aligner (MGA) program, which was developed by collaborators in Germany to dramatically speed up signature development.

In fall 2002, three of the team's summer students who had graduated—Clinton Torres, Jason Smith, and Lam—were hired to complete the KPATH system. Meanwhile, Kuczmarski and Shea Gardner handled the constant demands for new pathogen signatures using the original pipeline.

In December 2002, Livermore's smallpox and related signatures were evaluated by the CDC. A year earlier, Livermore had anticipated the need for smallpox assays and had developed candidate signatures. Just after KPATH signatures passed extensive testing in January 2003, the CDC requested that the team process several new smallpox and near-neighbor genomes in anticipation of world events.

In March 2003, the CDC requested assistance in analyzing the newly sequenced Severe Acute Respiratory Syndrome (SARS) virus. The Livermore team processed both a Canadian genome and a CDC genome and returned a set of signature candidates within three hours. The U.S. Army later tested these signatures, and the CDC is considering several for validation.

"SARS presents a special situation," says Slezak. "We don't really know what near-neighbor species are like, so it is very hard to be sure which signatures will work when near-neighbor viruses are eventually discovered. At this time, 77 percent of the genome appears to be highly unique, but this clearly will not be the case once other related organisms are discovered and sequenced. However, the automated KPATH approach will be capable of capitalizing on new data, and within 30 minutes we should be able to know which regions of SARS still appear to be unique and therefore which signatures will continue to work."

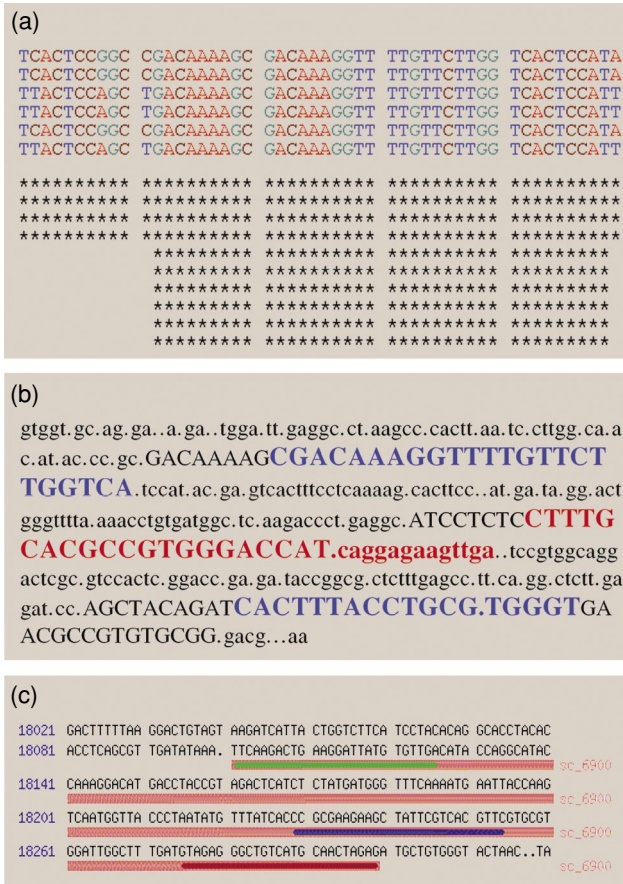
In June 2003, an unexpected outbreak of monkeypox in the U.S. occurred as a result of exotic animals being imported from Africa as household pets. Ironically, although the team had developed candidate signatures of monkeypox in 2002, the CDC had not tested these because monkeypox had never occurred in the Western Hemisphere. Following the U.S. outbreak, the CDC supplied Livermore with sequenced monkeypox from both a human and a prairie dog. The genomes, which turned out to be identical, were used by the Livermore team to refine the monkeypox signatures.

"These naturally occurring situations provide us with real-life training experiences for rapid emergency response," says Slezak. "It is a distinct honor that our team has earned the privilege of being the bioinformatics team assisting the CDC on major pathogen emergencies."

“In October 2000,” says Slezak, “we began building a preliminary pipeline based on this approach with funding obtained from the Laboratory Directed Research and Development Program. In May 2002, we were funded by DOE to build the current KPATH pipeline. We continued to use the first pipeline until about January 2003, when KPATH was shown to be functionally equivalent and much faster.”

New Features on the Horizon

Protein signatures are commonly used in diagnostic kits, such as commercially available home-use pregnancy tests. Slezak notes that the sequence of amino acids that make up a protein tends to be conserved (unchanged) because altering the protein sequence is likely to change the protein's shape, which in turn would alter its function. Using this approach, the team has found conserved and unique signature regions in the glycoprotein of the West Nile virus (an RNA virus) and has mapped



Another ongoing task for the team is building and maintaining relationships with partners in various agencies and universities. Slezak explains, “Much of the data we need are not in the public domain.”

KPATH uses computers and efficient algorithms to compare genomes and identify those portions that are unique to a particular pathogen or family of related pathogens. (a) A small part of the genome (50 DNA bases out of 8,000) of six strains of a pathogen are aligned and arranged in five columns for comparison. (The letters T, C, A, and G represent thymine, cytosine, adenine, and guanine, the four nucleic acids that make up all DNA.) The stars are a rough visual indicator of the degree of similarity or "consensus" among the six strains in each column. (b) A "consensus genome" is derived from those regions that are common to the six strains, indicated by the upper case letters. (c) The consensus region is mined to obtain a unique signature of the pathogen, which corresponds to the colored sequence.

—*Arnie Heller*

Key Words: BASIS, bioinformatics, bioterrorism, BioWatch, Centers for Disease Control and Prevention (CDC), DNA, foot-and-mouth disease, Human Genome Project, KPATH, microbe, pathogen, protein signature, RNA, Severe Acute Respiratory Syndrome (SARS), smallpox.

**For further information contact
Tom Slezak (925) 422-5746
(slezak1@ltnl.gov).**